Statistique et analyse des données

S. Joly
G. Le Calve

Étude des puissances d'une distance

Statistique et analyse des données, tome 11, n° 3 (1986), p. 30-50
<http://www.numdam.org/item?id=SAD_1986__11_3_30_0>

Statistique et Analyse des Données
1986 - vol.11 n°3 pp.30-50

ETUDE DES PUISSANCES D'UNE DISTANCE

S. JOLY, G. LE CALVE

Laboratoire d'Analyse des Données
Université de RENNES 2
6 Avenue Gaston Berger
35043 RENNES CEDEX

Résumé : *Le but de cet article est de rassembler quelques anciens et nouveaux résultats sur les propriétés des puissances d'une distance. On insiste sur le problème de l'existence d'une représentation euclidienne pour la puissance d'une distance. On établit une généralisation du lemme de Schur sur le produit d'Hadamard de deux matrices, le dernier paragraphe étant consacré à des applications, en analyse des données, de ce résultat.*

Abstract : *The aim of this paper is to gather some olds and news results on the properties of the power of a distance. The problem of the existence of an euclidean representation for such a power distance is emphasize. A generalisation of the Schur's lemma on the Hadamard product of matrices is given, the last section being devoted to applications, in data analysis, of this results.*

Mots clés : *Similarités, dissimilarités, représentations euclidiennes, formes quadratiques S D P.*

Indices de classification STMA : *03-060, 06-010, 06-020.*

Manuscrit reçu le 2 octobre 1986
Révisé le 7 avril 1987

0 - INTRODUCTION

Etudier les propriétés des puissances d'une distance est une idée qui vient assez naturellement à l'esprit, que ce soit à partir de la structure des métriques de Minkowski (D_k), ou à partir des transformations monotones de distance, dont les puissances sont un cas particulier intéressant.

Dans un premier paragraphe, on tente de faire le point sur l'état de la question, rassemblant des résultats, anciens pour certains, mais souvent mal connus, le plus typique étant sans doute le résultat de Schoenberg sur les puissances d'une distance euclidienne.

Le second paragraphe est tout entier consacré à une généralisation du lemme de Schur sur l'aspect semi défini positif du produit d'Hadamard de 2 matrices.

Cette généralisation est utilisée dans le paragraphe 3, tant pour donner des démonstrations nouvelles, dans une optique d'analyse des données, de résultats connus, que pour donner des propositions nouvelles.

Dans ce qui suit, le terme de dissimilarité fera toujours référence à une matrice carrée à termes réels positifs, symétrique, et à diagonale nulle ; le terme de distance s'entendra comme une dissimilarité dont les éléments vérifient l'inégalité triangulaire et l'axiome de séparabilité : $D_{ij} = 0 \iff i = j$.

Une dissimilarité $D = (D_{ij})$ sera dite euclidienne si et seulement si il existe n points $M_1 \ldots M_n$ d'un espace euclidien vérifiant

$$||M_i M_j|| = D_{ij}$$

1 - QUELQUES RESULTATS ET PROPOSITIONS

La plupart des résultats ci-dessous sont plus ou moins connus. Il nous paraît cependant utile de les réunir et les complémenter par des propositions nouvelles.

Résultat 1 (cf. par ex [15])

> Si $D = (D_{ij})$ est une distance, alors
>
> $\forall_0 \leqslant \alpha \leqslant 1 \qquad D^\alpha = (D^\alpha_{ij})$ est une distance.

Le résultat, immédiat, découle de ce que

$$0 \leqslant \alpha \leqslant 1 \qquad (a+b)^\alpha \leqslant a^\alpha + b^\alpha$$

Comme corollaire de ce résultat nous tirons la proposition:

Proposition 1

> Pour toute dissimilarité $D = (D_{ij})$ il existe un réel positif $p \geqslant 0$ dépendant de D, et àppelé "index de distance" tel que $D^\alpha = (D^\alpha_{ij})$ soit une distance pour $\alpha \leqslant p$ et ne le soit pas pour $\alpha > p$.

Le résultat précédent implique que s'il existe $p \in R^+$ tel que D^p soit une distance, alors $\forall q < p$, D^q est une distance. Comme limite D^α est une distance ($D^0_{ii} = 0$, $D^0_{ij} = 1$ pour $i \neq j$), on est assuré de $\overset{\alpha \to 0}{}$ l'existence d'un tel p.

On remarquera qu'une dissimilarité est une distance si et seulement si son index est supérieur ou égal à 1. On obtient également une caractérisation des ultramétriques.

Résultat 2 BROSSIER, LE CALVE [3]

> D est une ultramétrique si et seulement si son index de distance est infini.

Si D est une ultramétrique l'inégalité

$$D_{ij} \leqslant \text{Max}(D_{ik}, D_{jk}) \quad \text{entraîne évidemment que}$$

$$D^\alpha_{ij} \leqslant \text{Max}(D^\alpha_{ik}, D^\alpha_{jk}) \quad \forall \alpha \ .$$

D^α est donc une ultramétrique, et a fortiori une distance, ce qui implique que l'index d'une ultramétrique est infini.

Réciproquement, en supposant que D_{ij} soit la plus grande des trois distances D_{ik}, D_{jk}, D_{ij}, l'inégalité

$$D_{ij}^{\alpha} \leqslant D_{ik}^{\alpha} + D_{jk}^{\alpha} \quad \forall\alpha \quad \text{équivalente à}$$

$$1 \leqslant \left(\frac{D_{ik}}{D_{ij}}\right)^{\alpha} + \left(\frac{D_{jk}}{D_{ij}}\right)^{\alpha}$$

n'est possible pour tout α que si $D_{ij} = D_{ik}$ ou D_{jk} ce qui entraîne que D est une ultramétrique.

Remarquons de plus que, toute distance ultramétrique étant euclidienne (cf [9]), l'ultramétrique peut également se caractériser comme la seule distance dont toutes les puissances soient euclidiennes.

Cette constatation amène naturellement la question suivante : Toute distance admet-elle une puissance qui soit euclidienne ? Nous verrons plus loin qu'il est possible d'apporter une réponse positive à cette question et donc de créer un index d'Euclide. Depuis la première moitié du siècle une conjecture, cf. par exemple [15], assurait que cet index était égal à la moitié de l'index de distance, ce qui se formulait

"La racine carrée d'une distance est toujours euclidienne".

En 1936 Blumental apporte sa contribution à cette conjecture en démontrant le résultat suivant :

Résultat 3 : Propriétés des quatre points

> Sur quatre points la racine carrée d'une distance est toujours euclidienne.

Propriété à rapprocher de la propriété des trois points, affirmant que sur trois points toute distance est euclidienne.

Il est possible d'apporter une réponse définitive à cette question.

Proposition 2 FICHET - LE CALVE (non publié)

> Sur plus de quatre points la racine carrée d'une distance n'est pas toujours une distance euclidienne.

Il suffit de considérer le contre exemple suivant :

Soit D =

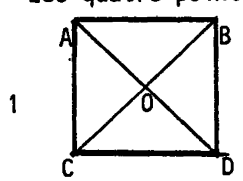
	A	B	C	D	E
A	0	1	2	1	1
B	1	0	1	2	1
C	2	1	0	1	2
D	1	2	1	0	1
E	1	1	2	1	0

D est une distance, $D^{\frac{1}{2}}$ aussi (ne serait-ce qu'en appliquant le résultat 1), mais $D^{\frac{1}{2}}$ n'est pas euclidienne.

$D^{\frac{1}{2}}$ =

	A	B	C	D	E
A	0	1	$\sqrt{2}$	1	1
B	1	0	1	$\sqrt{2}$	1
C	$\sqrt{2}$	1	0	1	$\sqrt{2}$
D	1	$\sqrt{2}$	1	0	1
E	1	1	$\sqrt{2}$	1	0

Les quatre points ABCD forment une figure plane, un carré de côté 1 :



E à égale distance de A, B, D ne peut donc se situer que sur l'orthogonale au plan passant par O. Auquel cas on aurait EA = EB = EC = ED, ce qui n'est pas le cas.

Le résultat 3 peut cependant s'étendre à plus de quatre points, mais il faut particulariser le type de distance.

On a déjà vu que c'était le cas pour la distance ultramétrique, dont toutes les puissances sont euclidiennes, et on en verra d'autres plus loin.

1.1 - Distances quadrangulaires

Une généralisation intéressante de la distance ultramétrique est la distance quadrangulaire (encore appelée : distance additive d'arbres, distance arborée, distance quadripolaire, etc ...).

Définition

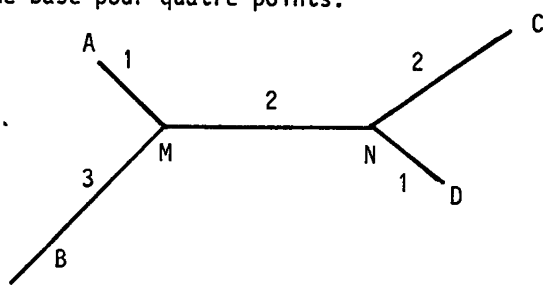
Une distance est dite quadrangulaire si et seulement si elle vérifie l'inégalité :

$$\forall i,j,k,\ell : \quad D_{ij} + D_{k\ell} \leqslant \mathrm{Max}(D_{ik} + D_{j\ell},\ D_{i\ell} + D_{jk})$$

Cette distance admet comme représentation un graphe planaire que l'on appelle arbre additif [13], [15] et dont voici un exemple, constituant d'ailleurs la configuration de base pour quatre points.

D =

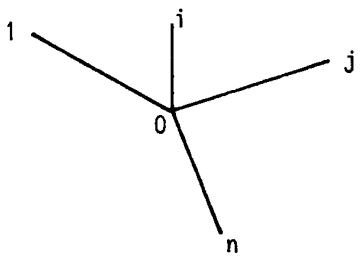
	A	B	C	D
A	0	4	5	4
B	4	0	7	6
C	5	7	0	3
D	4	6	3	0



La convention de lecture disant que la distance entre deux points est égale à la longueur de l'unique chemin qui les relie :

$$D_{AC} = |AM| + |MN| + |NC|$$

Le cas particulier où l'inégalité quadrangulaire est une égalité est connu sous le nom de "distance à centre" cf. [10]. Pour n points, elle admet la représentation en étoile ci-dessous avec la même convention de lecture que précédemment .



On a le résultat suivant :

Résultat 4 LE CALVE [10]

La racine carrée d'une quadrangulaire est euclidienne et de dimension maximum.

Rappelons que si D est euclidienne sa dimension est égale à la dimension de l'espace afine engendrée par les points de l'une quelconque de ses images euclidiennes.

Par dimension maximum, nous entendons que pour n points sa dimension est n-1.

La démonstration repose sur 3 remarques importantes en elles-mêmes (cf. [3]).

Remarque 1 :

La première est que l'additivité en D se traduit par l'orthogonalité en $D^{\frac{1}{2}}$:

$$D_{AB} = D_{BC} + D_{AC}$$

est le théorème de Pythagore en $D^{\frac{1}{2}}$

$$(D_{AB}^{\frac{1}{2}})^2 = (D_{BC}^{\frac{1}{2}})^2 + (D_{AC}^{\frac{1}{2}})^2$$

Remarque 2 :

La seconde concerne l'inégalité qui sert à définir la distance quadrangulaire. De manière analogue à ce qui est fait pour l'ultramétrique, il est possible de montrer que l'inégalité est équivalente (à une permutation près sur les indices) à

$$D_{ij} + D_{k\ell} = D_{i\ell} + D_{kj} \geqslant D_{ik} + D_{\ell j}$$

Remarque 3 :

A,B,C,D étant quatre points d'un espace euclidien, la formule du quadrangle dit que :

$$2(\vec{AB}, \vec{CD}) = AD^2 + BC^2 - AC^2 - BD^2$$

(si on identifie dans cette formule B et C on retrouve la formule bien connue du triangle).

Il est clair que dans la métrique $D^{\frac{1}{2}}$ la formule devient :

$$2(\vec{AB} \cdot \vec{CD})_{D^{\frac{1}{2}}} = AD + BC - AC - BD$$

et donc, d'après la remarque 2 que, dans une quadrangulaire tout quadrilatère possède deux côtés opposés orthogonaux. Ceci entraîne la dimension maximum.

1.2 - Distances euclidiennes

Après les ultramétriques et les quadrangulaires, il est logique de s'interroger sur les euclidiennes elles-mêmes.

Il appartient à Schoenberg d'avoir apporté en 1937, une réponse à ce problème en démontrant :

Résultat 5 SCHOENBERG [15]

> Soit D une distance euclidienne, alors $\forall \alpha < 1$
> D^{α} est une distance euclidienne de dimension maximum.

Ce résultat, bien qu'ancien déjà, semble peu connu. Schoenberg qui se situe dans un tout autre contexte, en donne une démonstration analytique, utilisant la notion de "fonction semi définie positive" qu'il introduit et étudie.

La démonstration que nous allons donner, outre qu'elle nous semble plus simple, repose sur une généralisation du lemme de Schur à propos du produit d'Hadamard de deux matrices semi définies positives.

2 - UNE GENERALISATION DU LEMME DE SCHUR

Rappel 1

Pour montrer qu'une distance est euclidienne on utilise la "forme associée à D".

Soit D une matrice $(n \times n)$ carrée symétrique réelle à diagonale nulle et M un point quelconque mais fixé. La forme associée à D au point M est la matrice

$$W^{M}(D)_{ij} = \frac{1}{2} (D_{Mi} + D_{Mj} - D_{ij})$$

Schoenberg (1935 [14]) montre que une CNS pour que D soit euclidienne est que $W^{M}(D^2)$ soit semi définie positive (en abrégé SDP), c'est-à-dire :

$$\forall X \in R^n \qquad \sum_{ij} W^M(D^2)_{ij}\, X_i X_j \geqslant 0$$

La dimension de l'espace de représentation est égale au rang de la matrice $W^M(D^2)$.

A propos de la paternité, douteuse, de ce résultat, John S. Lew [11] donne les précisions suivantes "résultat noté par Gauss (1831) dans le cas de trois dimensions, et prouvé par Dirichlet (1850) dans le cas de n dimension et exposé sous cette forme par Minkowski (1891), de nouveau mentionné par Frechet (1935) pour des espaces de Hilbert et appliqué par Schoenberg (1935, 1937) à la géométrie métrique".

Plus tard Torgerson (1958) propose de prendre pour M le centre de gravité de la figure engendrée, c'est-à-dire le point G défini par :

$$D^2_{Gi} = \frac{1}{n} \sum_j D^2_{ij} - \frac{1}{2n^2} \sum_{ij} D^2_{ij}$$

Rappel 2 Produit d'Hadamard

A et B étant deux matrices $n \times p$ leur produit d'Hadamard est la matrice $C = A * B$ définie par $C_{ij} = A_{ij} \cdot B_{ij}$.

Il s'agit du produit terme à terme des deux matrices. En 1911 Schur démontre le résultat suivant.

Lemme de Schur [16]

> A et B étant deux matrices carrées symétriques SDP leur produit d'Hadamard est SDP.

On sait d'autre part que la somme de deux matrices SDP est elle-même SDP. On est donc assuré que si A est SDP, il en est de même pour $\sum_{r=0}^{p} C_r A^{*r}$, $C_r \geqslant 0$; A^{*r} désignant la puissance $r^{ième}$ de A au sens d'Hadamard.

Il est donc possible d'étudier les séries entières, à coefficients positifs, de matrices SDP. On a le résultat :

Proposition 3 lemme de Schur généralisé

Soit $t \to f(t)$ une fonction réelle, développable en série entière à coefficients positifs, de rayon de convergence R et continue en $t = R$:

$$f(t) = \sum_{r=0}^{\infty} C_r t^r \qquad \forall t < R$$

Soit $A = (a_{ij})$, $a_{ij} \leqslant R$ une matrice carrée symétrique $(n \times n)$ SDP. Alors

i) La matrice $B = (B_{ij})$ où $B_{ij} = f(a_{ij})$ est SDP

ii) La CNS pour que B soit DP est que A vérifie la condition (C) (DP : definie positive).

Condition (C)

A, matrice SDP, vérifie la condition (C) si et seulement si $\forall X \in R^n$ il existe un entier r, pouvant dépendre de X, tel que :

$$\sum_{ij} X_i a_{ij}^r X_j = X^T A^{*r} X > 0$$

Démonstration

Elle se fait en considérant deux cas :

1er cas : si $a_{ij} < R \quad \forall ij$ alors $B = f(A) = \Sigma C_r A^{*r}$,

$$X^T B X = \sum_r C_r X^T A^{*r} X \geqslant 0$$

puisque, A étant SDP, il en est de même de A^{*r} (lemme de Schur).

D'autre part B SDP non DP est équivalent à :

il existe $X \in R^n$ tel que $X^t B X = 0$ et donc
il existe X tel que $X^t A^{*r} X = 0 \quad \forall r$

ce qui démontre le lemme dans le cas 1.

2ème cas : il existe des couples ij tels que $A_{ij} = R$

Posons :

$$B_\varepsilon = (B_{ij})_\varepsilon \quad \text{avec} \quad (B_{ij})_\varepsilon = f[a_{ij}(1-\varepsilon)]$$

la matrice B_ε est SDP d'après le cas 1.

Comme la continuité de f(t) en t = R implique :

$$\lim_{\varepsilon \to 0} X^T B_\varepsilon X = X^T B X \quad , \quad B \text{ est SDP}$$

Pour tout X posons

$$g_X(\varepsilon) = X^T B_\varepsilon X = \sum_r C_r (1-\varepsilon)^r X^T A^{*r} X$$

La condition C est alors équivalente à $g_X(\varepsilon)$ et strictement décroissante. Il en découle

$$X^T B X = \lim_{\varepsilon \to 0} X^T B_\varepsilon X > X^T B_\varepsilon X \geqslant 0$$

et donc B est DP si et seulement si la condition (C) est vérifiée.

REMARQUES SUR LA CONDITION C

Remarque 1

La condition (C) porte sur A et non sur f. Si donc g et f sont deux fonctions vérifiant les conditions du lemme g(A) est DP si et seulement si f(A) est DP.

Remarque 2

Dans la pratique la fonction f pour laquelle on étudie la positivité de f(A), appartient souvent à une famille plus vaste f_r. La remarque précédente ramène alors la condition (C) sur A à la même condition sur l'ensemble des $\{f_r\}$ ou même sur $\lim_r (f_r)$.

C'est ainsi que pour la famille $f_r(A) = A^{*r}$ on a :

Remarque 3

- Si $\lim_{r \to \infty} (A^{*r}) = I$, A vérifie la condition (C). En effet dans ce cas :

$$\forall X \quad \lim_{r \to \infty} (X^T A^{*r} X) = X^t X = \|X\|^2 > 0$$

et donc il n'existe pas de X tel que $A^{*r} X = 0 \ \forall r$. Ce résultat peut s'étendre

- Si il existe $k > 0$ tel que $\lim_{r\to\infty} k^r A^{*r} = I$, A vérifie la condition (C). Il suffit de poser $B = kA$; $\lim_{r\to\infty} B^{*r}$ est DP et les valeurs propres de A sont $\frac{1}{k}$ celles de B.

- Si il existe une matrice M de rang 1 telle que $\lim_{r\to\infty} (M^{*r} * A^{*r}) = I$, alors A vérifie la condition (C).

En effet $\forall X$ il existe Y tel que $X^T AX = Y^T M * AY$, ce qui entraîne le résultat en appliquant la remarque 3, premier point, à la matrice $B = M * A$.

Proposition 4

Soit S une matrice carrée symétrique à diagonale positive, vérifiant $S_{ii} > |S_{ij}|$ pour tout ij. Alors S vérifie la condition (C).

Posons :

$$R_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}\ S_{jj}}}$$

La matrice R a tous ses éléments diagonaux égaux à 1 et ses éléments non diagonaux strictement inférieurs à 1. Elle vérifie $\lim_{r\to\infty} [R^{*r}] = I$. En posant $M_{ij} = \sqrt{S_{ii}\ S_{jj}}$, S peut s'écrire sous la forme $M * R$ et S vérifie la condition (C) d'après le point trois de la remarque précédente.

Il en découle que toute matrice de similarité vérifie la condition (C) ainsi que toute matrice de corrélation (sous réserve que $|R_{ij}| \neq I$ pour $i \neq j$).

3 - APPLICATIONS DU LEMME DE SCHUR GENERALISE

Les résultats du paragraphe 2 trouvent un certain nombre d'applications dans le domaine qui nous occupe.

Le fait d'être euclidienne pour une distance se ramenant à la (semi) positivité d'une forme bilinéaire et les théorèmes en analyse des données sur le sujet n'étant pas légion nous utiliserons le lemme de Schur généralisé, soit pour donner de nouveaux résultats, soit pour donner de nouvelles démonstrations.

Proposition 5

Soit S une matrice de similarité SDP. Les matrices de terme général

$$A_{ij} = \frac{1}{K - S_{ij}} \quad K > S_{ii} \;, \quad B_{ij} = 1 - \sqrt{1 - S_{ij}} \;; \quad C_{ij} = 1 - (1 - S_{ij})^{\alpha} \quad 0 \leqslant \alpha \leqslant 1$$

$$E_{ij} = \frac{1}{K - S_{ij}^{\alpha}} \quad \begin{matrix} \alpha \geqslant 1 \\ K \geqslant S_{ii} \end{matrix} \;, \quad F_{ij} = \exp(S_{ij}) \;; \quad G_{ij} = \mathrm{Log}\left[\frac{K + S_{ij}}{K - S_{ij}}\right] \quad K > S_{ii}$$

$H = \text{Arc sinus } [S_{ij}]$

sont DP.

Il suffit de remarquer que chacune des fonctions de S_{ij} entrant dans la définition des matrices considérées est développable en série entière à coefficients positifs, et que S vérifie la condition (C).

Gower prouve dans [7] que A est SDP et ZEGERS dans [19] que B est SDP.

Remarque

- D étant une dissimilarité à valeur dans [0,1], il est habituel de lui associer une similarité S comme fonction décroissante de D dans [0,1].

Si $S_{ij} = 1 - D_{ij}$ est la plus répandue, il n'en est pas moins vrai qu'une fonction de type 1/D est acceptable sous certaines conditions ; certains auteurs proposent comme indice de similarité

$$T_{ij} = \frac{1}{1 + D_{ij}} - 1.$$

En remarquant que ces deux transformations sont liées par

$$T_{ij} = \frac{S_{ij}}{2 - S_{ij}} \;,$$

il en découle immédiatement que si S est SDP, T est DP (la réciproque étant fausse en général).

- Les résultats précédents sont également intéressants dans le domaine des indices de similarité sur des variables dichotomiques, que l'on peut interpréter en terme de (présence / absence). Il en existe un grand nombre et on trouvera une étude détaillée dans [6] et dans [7].

Ils sont en général construits de la façon suivante : i et j étant deux caractères, on définit :

n_{ij} = nombre de sujets présentant simultanément les caractères i et j

$\tilde{n}_{ij}$ = nombre de sujets ne présentant aucun des caractères i et j

q_{ij} = nombre de sujets ne présentant qu'un seul des caractères i ou j

n = nombre total de sujets.

Résulat 6 FICHET - LE CALVE [6]

Les racines carrées des indices de similarité S_{ij} définis par

n_{ij} / n	(Russel et Rao)
$(n_{ij} + \tilde{n}_{ij}) / n$	(Kendall, Sokal et Michener)
$n_{ij} / (n_{ij} + q_{ij})$	(Jaccard)
$2n_{ij} / (2n_{ij} + q_{ij})$	(Czenakowski, Dice)
$n_{ij} / (n_{ij} + 2q_{ij})$	(Sokal, Sneath, Anderberg)
$(n_{ij} + \tilde{n}_{ij}) / (n + q_{ij})$	(Rogers et Tanimoto)
$(n_{ij} - q_{ij} + \tilde{n}_{ij}) / n$	(Hamman)

sont DP.

Les racines carrées des distances qui leur sont associées par $D_{ij} = 1 - S_{ij}$ sont euclidiennes de dimension maximum.

Il suffit de remarquer qu'ils sont des fonctions homographiques de n_{ij}/n ou de $(n_{ij} + \tilde{n}_{ij}) / n$ et d'appliquer la proposition 5.

NOUVELLE DEMONSTRATION DU RESULTAT DE SCHOENBERG

Comme autre application nous pouvons donner une autre démonstration du résultat de Schoenberg qui dit que

$$D \text{ euclidienne} \Rightarrow D^{\alpha} \text{ euclidienne} \qquad 0 \leqslant \forall \alpha \leqslant 1$$

et de dimension maximum.

Démonstration

Elle se fait en considérant deux cas :

1er cas : il existe un point O tel que $D_{oi} = 1 \quad \forall i$. La forme associée à D^2 au point O définie par

$$2\mathring{W}(D^2)_{ij} = D^2_{oi} + D^2_{oj} - D^2_{ij}$$

s'écrit :

$$2\mathring{W}(D^2)_{ij} = 2 - D^2_{ij}$$

et, de la même façon :

$$2\mathring{W}(D^{2\alpha})_{ij} = 2 - D^{2\alpha}_{ij}$$

Il en découle que

$$2\mathring{W}(D^{2\alpha})_{ij} = 2 - 2^{\alpha}[1 - \mathring{W}(D^2)]^{\alpha}$$

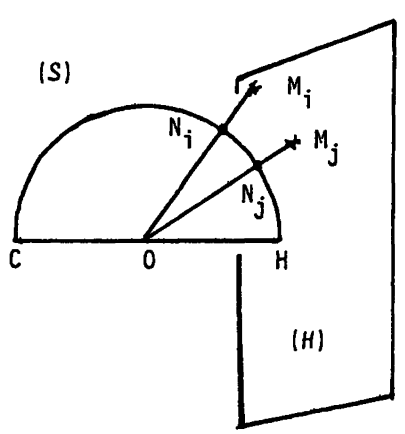
$\mathring{W}(D^{2\alpha})$ est donc une fonction de $\mathring{W}(D^2)$ vérifiant les conditions de la proposition 3 et donc SDP.

Comme, de plus, limite $\mathring{W}(D^{2\alpha})$ est DP (elle vaut 1 sur la diagonale et $\frac{1}{2}$ en dehors) $\mathring{W}(D^{2\alpha})$ est $\overset{\alpha \to o}{DP}$ d'après la remarque 2 de la proposition 3.

2ème cas : la figure associée à D n'a pas de point central. Soit C un point à la distance 2 du sous espace (H) engendré par les points de la figure, et soit H la projection de C (sur H). L'inversion de pôle C et de puissance 4 définie par

$$\overline{CM_i} \cdot \overline{CN_i} = 4, \quad \text{et donc} \quad CH^2 = 4,$$

expédie le sous espace (H) sur la sphère (S) de diamètre CH. Soit O le milieu de CH.



La distance D_{ij} est ainsi transformée en une distance

$$\delta_{ij} = \frac{D_{ij}}{D_{ci}\, D_{cj}}$$

(les triangles CM_iM_j et CN_iN_j sont semblables) et donc

$$W^c(\delta)_{ij} = \frac{1}{D_{ci}\ D_{cj}}\ W^c(D)$$

La distance δ vérifie la condition du cas 1 puisque $\delta_{oi} = 1 \quad \forall i$ donc δ^α est euclidienne et de dimension maximum.

Donc

$$W^c(\delta^\alpha) \quad \text{est DP}$$

or

$$W^c(\delta^\alpha) = \frac{1}{D^\alpha_{ci}\ D^\alpha_{cj}}\ W^c(D^\alpha)$$

$W^c(D^\alpha)$ est donc DP d'après le lemme de Schur, il en découle que D^α est euclidienne de dimension maximum.

REMARQUE SUR LE PROBLEME DE LA CONSTANTE ADDITIVE

Sous ce nom on désigne en général le problème suivant :

Problème A : D étant une dissimilarité (D_{ij}) trouver c minimum tel que la dissimilarité définie par $(D_{ij} + c)_{i \neq j}$ soit euclidienne.

En 1983, CAILLIEZ cf. [4] a résolu le problème suivant :

Problème B : D étant une dissimilarité donnée, trouver c^* minimum tel que $\forall c > c^*$ la dissimilarité définie par $(D_{ij} + c^*)_{i \neq j}$ soit euclidienne.

Les résultas précédents nous permettent de montrer que

Proposition 7

Si (D_{ij}) est une dissimilarité euclidienne alors, $\forall c \geqslant 0 \quad (D_{ij} + c)_{i \neq j}$ est euclidienne. Il en découle que les Problèmes A et B sont équivalents.

Par commodité d'écriture nous notons D_c la nouvelle dissimilarité.

Démonstration

Pour montrer que D_c est euclidienne il faut montrer que

$$X^T W^M (D_c^2) X \geqslant 0 \qquad \forall X \text{ de } R^n$$

or

$$X^T W^M(D_c^2)X = X' W^M(D^2)X + 2cX^T W^M(D)X + \frac{c^2}{2} X^T AX$$

où A est la matrice valant 1 sur la diagonale et $\frac{1}{2}$ en dehors. A est DP et $W^M(D)$ l'étant également en raison de la proposition 6, $X^T W^M(D_c^2)X$ est donc une fonction réelle strictement croissante de c, donc :

$$W^M(D^2) \quad \text{SDP} \quad \Longrightarrow \quad W^M(D_c^2) \quad \text{DP}$$

INDEX D'EUCLIDE

Proposition 8

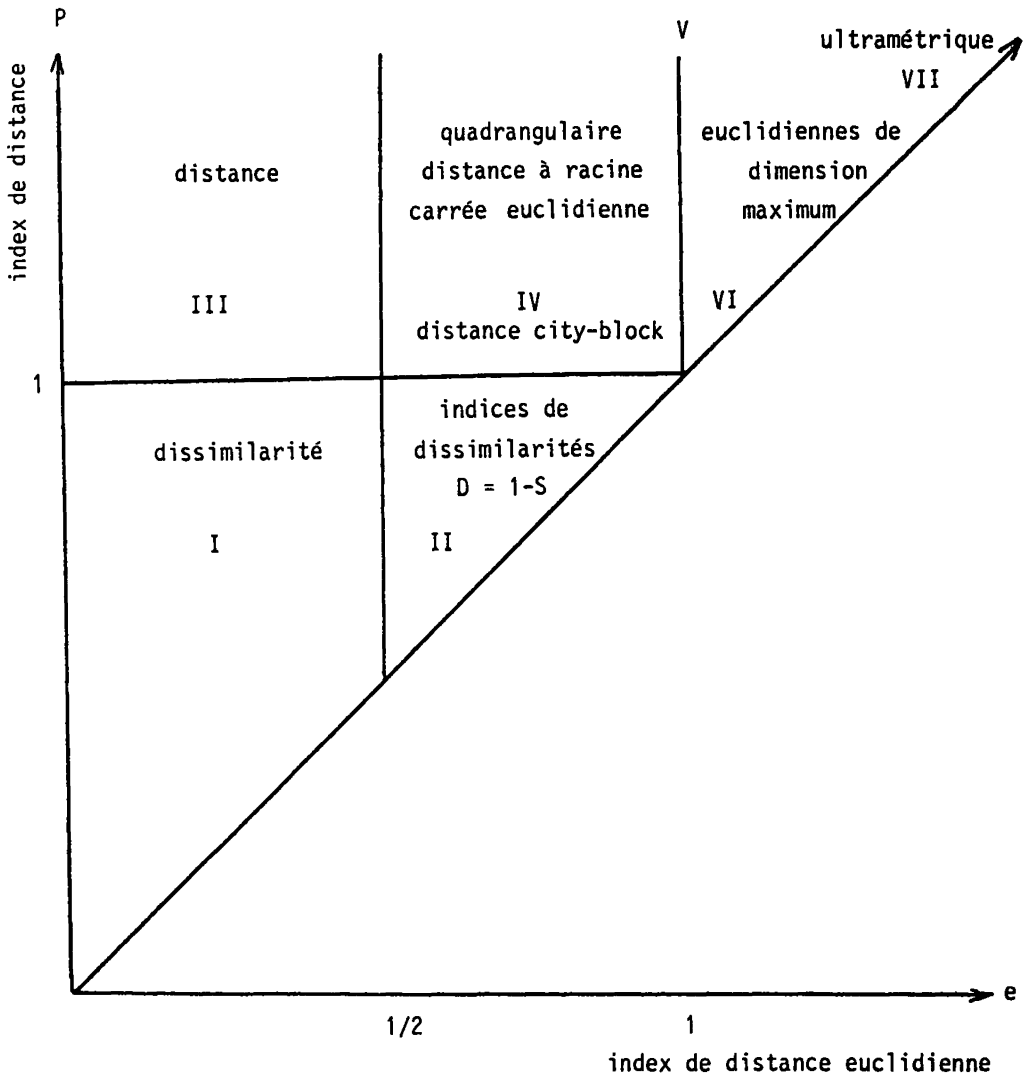
> Pour toute dissimilarité $D = (d_{ij})$ il existe un réel positif e, dépendant de D, appelé index d'Euclide, et tel que $D^\alpha = (d_{ij}^\alpha)$ soit une distance euclidienne pour tout $\alpha < e$ et ne soit pas euclidienne $\forall \alpha > e$.

On sait que si D^e est euclidienne, il en est de même de D^α $\forall \alpha < e$ d'après la proposition 6.

Comme limite D^α est une distance euclidienne, on est assuré de l'existence d'un tel e.
$\alpha \to 0$

ESSAI DE TABLEAU SYNOPTIQUE

En utilisant les propositions 1 et 8, on peut donner le graphique suivant, qui nous servira de conclusion, illustrant la "structure" de divers indices de dissimilarité.


P
V
ultramétrique
VII
index de distance
distance
quadrangulaire
distance à racine
carrée euclidienne
euclidiennes de
dimension
maximum
III
IV
distance city-block
VI
1
indices de
dissimilarités
D = 1-S
dissimilarité
I
II
e
1/2
1
index de distance euclidienne

ZONE I : $e < \frac{1}{2}$ et $P < 1$ on a affaire à des dissimilarités qui ne sont pas des distances.

ZONE II : $\frac{1}{2} < e \leqslant 1$, $P \leqslant 1$ on y trouve la plupart des indices sur variables dichotomiques. La distance engendrée n'est pas euclidienne mais la racine carrée est euclidienne.

ZONE III: $e < \frac{1}{2}$, $P \geqslant 1$ zone des distances dont la racine carrée n'est pas euclidienne.

ZONE IV : $\frac{1}{2} \leqslant e < 1$, $P \geqslant 1$ zone des distances dont la racine carrée est euclidienne mais qui qui ne sont pas euclidiennes. On y trouve en particulier les quadrangulaires (représentables par des arbres additifs) et les distances du type city-block (distance de la valeur absolue).

ZONE V : $e = 1$, $P \geqslant 1$ distances euclidiennes de dimension non maximum

ZONE VI : $e > 1$, $P > 1$ zone des distances euclidiennes de dimension maximum

ZONE VII : $p = e = \infty$ point des ultramétriques.

[11] LEW, "Some counter examples in multidimensional scaling", Journal of Mathematical Psychology, 1978, 17, pp. 247-254.

[12] PATRINOS, HAKINI, "The distance matrice and its tree realisation", Quaterly of applied mathematics, 1972, Vol. 30, n° 3.

[13] SATTAH, TVERSKY, "Additive similarity trees", 1977, Psychometrika 42-3.

[14] SCHOENBERG, "Remarks to Maurice Frechet's article", Annals of Mathematics, 1935, Vol. 36, pp. 724-732.

[15] SCHOENBERG, "On certain metric spaces arising from euclidean spaces by a change of metric and their imbedding in Hilbert space", Annals of Mathematics, 1937, Vol. 38, n° 4, pp. 787-793.

[16] SCHUR, "Bemerkungen zur Theorie der beschrankter Bilinearformen mit unendlich vielen Veränderlichen", Journal fur die reine und angewandte, MATHEMATIK, 1911, 140, pp. 1-28.

[17] TORGERSON, "Theory and methods of scaling", New York, Wiley, 1958.

[18] YOUNG, HOUSEHOLDER, "Discussion of a set of points in terms of their mutual distances", Psychometrika,3, 1938, pp.19-22.

[19] ZEGERS, "Two classes of elementwise transformations preserving the PSD nature of coefficients matrices", J. Classification, 1986, Vol. 3, n° 1, pp. 49-54.

BIBLIOGRAPHIE

[1] BLUMENTAL, "New theorems and methods in determinant theory", Duke Math. J., 1936, Vol. 2, pp. 396-404.

[2] BROSSIER, "Problèmes de représentation de données par des arbres", Thèse d'Etat, Université de RENNES 2, 1986.

[3] BROSSIER, LE CALVE, "Analyse des dissimilarités sous l'éclairage $\sqrt{D}$. Application à la recherche d'arbres additifs optimaux", in : Data analysis and informatics IV Diday et al. North Holland, 1985.

[4] CAILLEZ, "The analytical solution of the additive constant problem", Psychometrika, 1983, Vol. 48, n° 2, pp. 305-308.

[5] FRECHET, "Sur la définition axiomatique d'une classe d'espaces vectoriels distanciés applicables vectoriellement sur l'espace de Hilbert", Ann. Math. 1935, 36, pp. 705-718.

[6] FICHET, LE CALVE, "Structure géométrique des principaux indices de dissimilarité sur signes de présence absence", Statistiques et Analyse des données, 1984, Vol. 9, n° 3, pp. 11-44.

[7] GOWER, "Measures of similarity, dissimilarity and distance", in Encyclopedia of statistical sciences, 1985, Vol. 5, Eds S. Kotz, N.L. Johnson and C.B. Read. New York, John Wiley and sons, pp. 397-405.

[8] HADAMARD, "Leçons de géométrie élémentaire", 1931, Vol. 1, Paris.

[9] HOLMAN, "The relation between hierarchical and euclidean models for psychological distances", Psychometrika, 37, n° 4, 1972.

[10] LE CALVE, "Distances à centre", Statistiques et analyse des données", 1985, Vol. 10, n° 2, pp. 29-44.